Homework 4: More Principal Component Analysis

About

Due

Monday 2/18/19, 11:59 PM CST

Goal

This homework focuses on familiarizing you with low-rank approximations and multi-dimensional scaling. In addition, you will work with the CIFAR-10 dataset, a popular benchmark dataset for most classification algorithms.

Additionally, it is intended to provide practice with finding and using publicly available libraries, an essential skill when applying machine learning techniques.

Code and External Libraries

The assignment can be done using any language.

You may use external libraries to perform PCA, as well as to compute euclidean distances.

You are expected to write your own code for Principal Coordinate Analysis.

Problems

Total points: 100

CIFAR-10 is a dataset of 32x32 images in 10 categories, collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. It is often used to evaluate machine learning algorithms. You can download this dataset from https://www.cs.toronto.edu/~kriz/cifar.html (https://www.cs.toronto.edu/~kriz/cifar.html).

A. For each category, compute the mean image and the first 20 principal components. Plot the error resulting from representing the images of each category using the first 20 principal components against the category.
B. Compute the distances between mean images for each pair of classes. Use principal coordinate analysis to make a 2D map of the means of each categories. For this exercise,

compute distances by thinking of the images as vectors. (Follow procedure 7.2 on page 120 of the book)

C. Here is another measure of the similarity of two classes. For class A and class B, define E(A → B) to be the average error obtained by representing all the images of class A using the mean of class A and the first 20 principal components of class B. This should tell you something about the similarity of the classes. Now define the similarity between classes to be
(1/2)(E(A → B) + E(B → A)). Use principal coordinate analysis to make a 2D map of the classes. Compare this map to the map in the previous exercise – are they different? why?

Submission

Submission will be through gradescope (https://www.gradescope.com):

Your submission for this homework should include:

1. PDF report with the following to the **HW4 Report** portal:

   1. Page 1: **(10 points)** A plot of the mean images of each class
   2. Page 2: **(15 points)** A plot of the sum-squared error from representing a class with the first 20 principal components of that class, for each class.
   3. Page 3: **(25 points)** The 2D scatter plot obtained after performing principal coordinate analysis using euclidean distance.
   4. Page 4: **(25 points)** The 2D scatter plot obtained after performing principal coordinate analysis using the similarity metric in part C.

2. Code submission via gradescope to the **HW4 Code** portal:

   1. **(5 points)** Submit your code as a zipped file.
   2. **(10 points)** Submit a CSV containing the distance matrix between the mean images of each class as a CSV file named partb_distances.csv
   3. **(10 points)** Submit a CSV containing the distance matrix between the mean images of each class as a CSV file named partc_distances.csv